

-- ABSTRACT OF THE DISCLOSURE

92 A method and arrangement for speech recognition wherein a volume distance is determined between recognized words and the pauses lying between them. When the volume distance of a word is lower than a predetermined threshold, the word is evaluated as being incorrectly recognized, such that errors caused by unwanted noises are avoided. - -

REMARKS

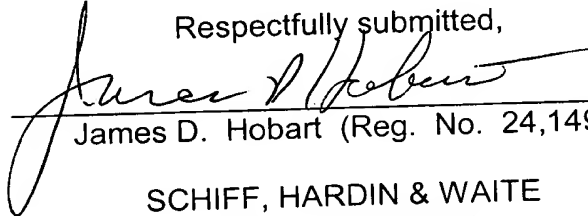
A substitute specification is provided herewith which makes editorial changes in order to conform to standard US practice. A marked-up copy of the specification is also provided reflecting the changes made.

In addition, the claims as filed have been cancelled and replaced by new claims that more clearly set forth the subject matter of Applicants' invention.

No new matter has been inserted in the application.

Applicants submit that this application is in proper condition for examination, which action is respectfully requested.

Respectfully submitted,



James D. Hobart (Reg. No. 24,149)

SCHIFF, HARDIN & WAITE  
Patent Department  
6600 Sears Tower  
233 South Wacker Drive  
Chicago, IL 60606  
Telephone: (312) 258-5790  
Attorneys for Applicant

Date: 12/22/2000

{Siemens AG}[Substitute Specification:]

~~{New application originally filed in German~~

~~26965-0214 (P-00,1701)~~

~~4999P02588US~~

~~USSN 642,452~~

~~Inventor: Bauer et al.~~

~~Translation / October 3, 2000 / 1696(911) / 3900 words~~

~~METHOD AND ARRANGEMENT)[- METHOD AND SYSTEM] FOR  
SPEECH RECOGNITION~~

**[BACKGROUND OF THE INVENTION]**

**Field of the Invention**

In general, the] {The} present invention is directed to {an method  
and an arrangement for speech recognition.} [speech recognition systems.

In particular, the present invention is directed to automatic detection of  
speech recognition errors.

**Discussion of the Related Art]**

Methods for automatic speech recognition are [often] utilized in  
speech recognition systems. Applications of speech recognition systems are,  
for example, dictating systems or automatically operating telephone  
exchanges.

{given} [It is especially critical in speech recognition that the  
correct expressions of the correct speaker are recognized. This is

problematical insofar as an ambient noise in which clear speech constituents are contained can be interpreted such by a speech recognition system as though they derived from the speaker of the speech actually to be recognized. In order to prevent a mix-up, a method is herewith disclosed for distinguishing the correct form the incorrect spoken language. In particular, the level of the speaker whose speech is to be recognized is usually clearly higher than speech from the unwanted noise, which usually comes from the background. The volume level of the speaker whose speech is to be recognized can thus be used to distinguish this from the background noise.

**Given]** previously known methods for the automatic recognition of speech recognition errors are frequently caused by unwanted noises. A distinction is made between two types of unwanted noises, namely the speech of another speaker that is in fact usually correctly recognized but that is not to be assigned to the voice signal of the actual speaker and a ~~{background}~~ **[back-ground]** noise not representing a voice signal such as, ~~{for example,}~~ breathing sounds, that is incorrectly recognized as speech. {  
}The unwanted noises represent a considerable source of error in the automatic recognition of speech.

In order to avoid such errors, speech recognition systems are trained to the speech of the individual speakers, so that the speech ~~{recognition}~~ **[recogni-tion]** system can determine whether the acoustic signal derives from the speaker or is a background noise. Speech recognition

systems having frequently changing speakers cannot be trained for every individual speaker. Given a speech recognition system integrated in a telephone system, thus, it is impossible to carry out a training phases lasting a number of minutes for every caller before the caller can speak his message, which often lasts only a fraction of a minute.

### **[SUMMARY OF THE INVENTION**

It is an] ~~{Advantageously, the}~~ object of the present invention ~~{is to enable a }~~ **[to provide a method for enabling]** recognition of speech wherein recognition errors produced by unwanted noises are reduced.

~~{This object is achieved according to the features of the independent patent claims. Developments of the invention also derive from the dependent claims:~~

~~For achieving the invention, a method for recognizing speech is specified wherein words and pauses in the speech are determined}~~ **[It is another object of the invention to provide a method for determining words and pauses ]**on the basis of word boundaries.

**[It is a further object of the invention to provide a method wherein an average silence volume can be determined during speech pauses.**

**It is an additional object of the invention to provide a method to determine average word volume.**

It is yet another object of the invention to provide a method to determine a difference between the average word volume and the average silence volume.

It is yet a further object of the invention to provide a method wherein speech is recognized when a difference between an average word volume and an average silence volume is greater than a predetermined threshold.

These and other objects of the invention will become apparent upon careful review of the following disclosure, which is to be read in conjunction with review of the accompanying drawing figures.

#### **BRIEF DESCRIPTION OF THE DRAWINGS**

Figure 1 shows a flowchart according to the present invention;

Figure 2 shows a part of a signal segment according to the present invention;

Figure 3 shows a circuit diagram of a telecommunication system according to the present invention.

#### **DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS**

Fig. 1 schematically shows a method for the automatic recognition of speech. This method is realized in practice by a computer program that works on a computer or a processing unit comprising an input for a voice signal.

The method is started in Step S1. In the following Step S2, a word of a speech signal S is analyzed. This analysis continues such that the acoustic voice signal which is usually present as a signal converted into an electrical signal, is segmented into words and pauses and the words are converted into text. The segmentation of the signal section is sometimes referred to as the Viterbi alignment method.]

An average silence volume ( $S_i$  level) is determined during the pauses. An average word volume ( $W_o$  level) for the words is also ~~{determined}~~ **[deter-mined]**. Further, a difference ( $\Delta$ ) between the average word volume ( $W_o$  level) and the average silence volume ( $S_i$  level) is also determined. Speech is recognized when the difference ( $\Delta$ ) between the average word volume ( $W_o$  level) and the average silence volume ( $S_i$  level) is greater than a ~~{predetermined}~~ **[predetermin-ed]** threshold ( $S$ ). Otherwise, a recognition of speech is not carried out in this range.

The difference  $\Delta$  forms a volume distance between the spoken words and the noises in the pauses. When the volume distance of a ~~{recognized}~~ **[recogn-ized]** word is too slight, it is interpreted as an incorrectly recognized word. A determination is thus made as to whether a word has a predetermined volume distance from the remaining noise level. ~~{The fact is thereby utilized that background}~~ **[Background]** noises that often lead to incorrect recognitions in traditional methods for automatic speech recognition are not as loud as the words spoken by the speaker. These background noises can simply be filtered out with ~~{the-}~~ **[using the method of the**

**present]** invention, regardless of whether they contain words or are noises that do not represent a voice signal.

The ~~{inventive}~~ method **[of the present invention]** can also be ~~{realized in a very simple way in}~~ **[used such]** that only the average volume need be determined over parts of the speech signal segment to be analyzed.

~~{What is understood as volume in the sense of the invention is}~~  
**[The term volume refers to]** any physical quantity that is ~~{approximately}~~  
**[approxima-tely]** proportional to the physical volume measured in decibels. Proportional quantities **[relating]** thereto are the energy of the acoustic signal or, ~~{respectively,}~~ **[respecti-vely,]** of ~~{a-ch}~~ **[an]** electrical signal ~~{and, in particular, the electrical quantities thereof}~~ such as, for example, ~~{the}~~ voltage or the current.

~~{It is especially critical in speech recognition that the correct expressions of the correct speaker are recognized. This is problematical insofar as an ambient noise in which clear speech constituents are contained can be interpreted such by a speech recognition system as though they derived from the speaker of the speech actually to be recognized. In order to prevent a mix-up, a method is herewith disclosed for distinguishing the correct form the incorrect spoken language. In particular, the level of the speaker whose speech is to be recognized is usually clearly higher than speech from the unwanted noise, which usually comes from the background. The volume level of the speaker whose speech is to be recognized can thus be used to distinguish this from the background noise.}~~

An arrangement for speech recognition is also recited for achieving the object, this comprising a processor unit that is configured such that

- a) words and pauses in the speech can be determined on the basis of word boundaries;
- b) an average silence volume ( $S_i$  level) can be determined during the pauses;
- c) an average word volume ( $W_o$  level) for the words can be determined;
- d) a difference ( $\Delta$ ) between the average word volume ( $W_o$  level) and the average silence volume ( $S_i$  level) can be determined;
- e) speech is recognized when the difference ( $\Delta$ ) between the average word volume ( $W_o$  level) and the average silence volume ( $S_i$  level) is greater than a predetermined threshold ( $S$ );
- f) otherwise, a recognition of the speech is not implemented.

This arrangement is especially suited for the implementation of the inventive method or one of the developments thereof explained above.

The invention is explained in greater detail below by way of example with reference to the appertaining drawings:

Shown in the drawings are:

Figure 1 a method for speech recognition shown schematically in a flowchart;

Figure 2 a diagram that represents a part of the signal segment;

Figure 3 a schematic block circuit diagram of a telecommunication system that works according to the inventive method.





present between the point-in-time zero and t1 or, ~~{respectively, following}~~  
[follow-ing] point-in-time t2 and the signal S represents a word between the  
points-in-time t1 and t2.

An average silence volume Si level is determined in Step S3. The  
average silence volume Si level is the chronological average of the volume of  
one or more pause segments P.

In Step S4, an average word volume Wo level defined. ~~{The~~  
average} [Average] word volume Wo level is the chronological average of the  
volume of an individual word segment W. I.e., a separate Wo level is  
calculated for each individual word.

In the following Step S5, a difference  $\Delta$  is calculated between the  
average word volume Wo level and the average silence volume Si level:

$$\Delta = \text{Wo-Level} - \text{Si-Level}$$

Subsequently, an interrogation is carried out in Step S6 to see whether the  
difference  $\Delta$  is lower than a threshold SW. The threshold SW represents  
"volume distance" (also see Fig. 2).

When this increase shows that the difference  $\Delta$  is smaller than the  
threshold SW, then this means that the volume distance between the average  
word volume Wo level and the average silence volume Si level is less than  
the predetermined threshold SW. The word whose volume distance between  
the average volume level Wo level and the average silence volume Si level is  
lower than the predetermined threshold SW is evaluated as having been  
incorrectly recognized, since the inventors of the present invention have

found that the unwanted noises are usually not as loud as the word signals to be evaluated or that, given a constant unwanted noise (noise in the line, loud background noise) where in no satisfactory speech recognition is possible, the volume distance between the average word volume and the average silence volume is extremely slight. When the acquired signal is converted into a text in both instances, it merely always results in an incorrect recognition. When the inquiry in Step S6 yields that the difference  $\Delta$  is lower than the threshold SW, ~~{then}~~ the program execution is branched to the Step S7 wherein an error elimination is implemented, this being explained later.

Subsequently, a check is carried out in Step S8 to see whether a further word is to be evaluated. When the result in Step S6 is that the difference  $\Delta$  is greater than the threshold SW, the program execution is directly branched onto an inquiry in Step S8.

A check is carried out with the inquiry in Step S8 to see whether a further word is yet to be analyzed and to be interpreted and, if the result in "yes", the program execution is branched back onto the Step S2; otherwise, the program is ended with Step S9.

~~{In the above-described exemplary embodiment, the acquired}~~

**[Acquired]** words are individually analyzed, converted into text and interpreted. This method is referred to as pace-keeping recognition. It is thereby expedient that the difference  $\Delta$  between the average word volume  $W_o$  level of a word W and the average silence volume  $S_i$  level o of the immediately preceding pause P is formed. However, it is also possible to

employ the average silence volume of the pause following the word W or to employ a silence volume averaged over the preceding or the following pause.

Instead of a pace-keeping recognition, a recognition combining several words can also be employed. A complete sentence is thereby usually respectively to be registered as signal segment and to be then analyzed of a piece (sentence-by-sentence recognition). Given such a sentence-by-~~{sentence}~~ **[sent-ence]** recognition, the silence volume can be averaged over all pauses P, whereby~~{, however,}~~ the average word volume is to be individually determined for each word W, so that the individual words can be evaluated as correctly or incorrectly recognized.

Dependent on the application, there are various versions in the error elimination in Step S7 which can be utilized individually or in ~~{combination. According to the first version, words}~~ **[combina-tion.**

**Words]** that have been evaluated as incorrectly recognized are not taken into consideration in the conversion into a text or~~{, respectively,}~~ are removed ~~{therefrom}~~ **[there-from]**.

According to ~~{the}~~ **[a]** second version of error elimination, a corresponding message is output to the user given a word deemed incorrectly recognized. The message can be output as an acoustic message (for example, "the last word was not correctly understood") or can be displayed as a graphic display. The former is expedient for speech recognition systems without display such as, for example, telecommunication system with automatic speech ~~{recognition}~~ **[recogni-tion]** and the second can be

meaningful, for example, given a dictating system. In dictating systems, a predetermined error character can be inserted at the corresponding location in the text as a graphic presentation, the user being prompted therewith to speak the word again, this then being automatically introduced at the location of the error character in the text. When the user does not wish to insert a word for this, he can actuate a correspondingly delete function for illuminating the error character.

According to a third version of the error illumination, the user can be prompted by a corresponding message to speak louder, so that the required volume distance is achieved. As a result thereof, an adaptation of the voice input to the acoustic conditions (noise level by the speaker) or, respectively, the conditions of the transmission (noise on the line) of the acoustic signal ensues. When a repeated prompt to speak louder does not lead to an improved recognition result, the user can also be prompted to create different acoustic conditions or, respectively, transmission conditions in that, for example, the user is requested to telephone from a different telephone set if the user is connected to the speech recognition system via a telephone.

According to a fourth version of the error elimination given a plurality of words successively evaluated as incorrectly recognized, this is evaluated as inadequate quality of the speech input and is indicated to the user with a corresponding message.



word volume of a plurality of acquired words and the average silence volume of a plurality of acquired pauses, and this global difference  $\Delta g_1$  can be employed as threshold SW, either directly or after the subtraction of a predetermined, constant amount. This is particularly ~~{advantageous}~~ **[useful]** in combination with the first version of the error elimination since background noises can also be filtered out as a result thereof, these being only slightly softer than the average word volume. The result thereof is that, given a speech input with high quality, the threshold below which the signals are evaluated as incorrectly recognized words is set higher than given a speech input with poorer quality. Preferably, a lower limit is provided for the threshold, so that this cannot be reduced to zero.

The height of the variable threshold can also be evaluated as quality factor of the speech input. When the variable threshold reaches its lower limit, then this means that the quality of the speech input is relatively poor, which can be correspondingly communicated to the user.

In the calculation of the global difference, all pauses and words that are spoken during a conversation with the speech recognition system are preferably taken into consideration.

Fig. 3 shows an exemplary embodiment of an apparatus for speech recognition. This apparatus is a telecommunication system 1 that is connected to the telephone network via a network line 2. The ~~{telecommunication}~~ **[telecommu-nication]** system 1 comprises a subscriber access control 3 with which ~~{telephone}~~ **[tele-phone]** subscribers calling from

the outside can be connected via an internal bus 4, a digital-to-audio processor 5 and local telephone lines 6 to a telephone terminal 7 or, respectively, to the user using the telephone terminal. The internal bus 4 is connected to an announcement unit 8 and to a voice unit 9. Announcements can be introduced onto the bus 4 and, thus, onto the telephone lines 2, 6 with the announcement unit 8. The telecommunication system is controlled by a microprocessor 10 that is connected to the digital-to-audio processor 5, to the announcement unit 8 and to the voice unit 9.

The voice unit 9 is composed of a speech analysis module 11, a volume measuring means 12 and a voice control 13.

The speech analysis module 11 carries out the analysis of the voice signal, whereby the voice signal is segmented into pauses and words and the words are converted into text. The speech analysis module conducts the individual parts (words W and pauses P) of the speech signal S to the volume measuring means 12 and forwards the converted text to the voice control 13. The volume measuring means determines the average volume ( $W_0$  level,  $S_i$  level) of the individual parts of the speech signal and forwards the corresponding values to the speech control 13. A check is carried out in the speech control 13 to see whether the individual words have been correctly recognized (Step S6 in Fig. 1), whereby filtering incorrectly recognized words is potentially undertaken in the speech control 13 (first version of the error elimination).



The filtered or unfiltered text is forwarded from the speech control 13 together with further data needed for the error elimination to the microprocessor 10 that evaluates the received text and the corresponding data.

One function of the microprocessor 10 is to automatically connect the incoming calls to the respective telephone terminals 7. This ensues by interpreting the text received from the speech control 13 and by enabling the respective output of the digital-to-audio processor 5.

When the received text cannot be interpreted or when an error elimination with announcements (second, third or fourth version) is necessary, then the announcement unit 8 is driven by the microprocessor to implement the corresponding announcement.

An automatic switching is thus integrated into the inventive telecommunication system, this being capable of automatically forwarding incoming telephone calls to the respective telephone terminals.

The {inventive} telecommunication system 1 also makes it possible that the users of the telephone terminals 7 control the telecommunication system 1 with their voice and, for example, speak the number to be selected instead of typing it on the keys.

All of these functions assume an optimally error-free speech recognition. As a result of the invention, errors due to background noises, whether as a result of a speech signal in the background or a noise that does

not represent a speech signal, can be avoided significantly better and in a simpler way than given traditional speech recognition systems.

{Patent Claims

~~1. Method for speech recognition, whereby~~

~~a) words and pauses in the speech are determined on the basis of word boundaries;~~

~~b) an average silence volume ( $S_i$  level) during the pauses is determined;~~

~~c) an average word volume ( $W_o$  level) for the words is determined;~~

~~d) a difference ( $\Delta$ ) is determined between the average word volume ( $W_o$  level) and the average silence volume ( $S_i$  level);~~

~~e) whereby speech is recognized when the difference ( $\Delta$ ) between the average word volume ( $W_o$  level) and the average silence volume ( $S_i$  level) is greater than a predetermined threshold ( $S$ );~~

~~f) otherwise, no recognition of the speech is implemented.~~

~~2. Method according to claim 1, whereby the average silence volume and the average word volume is measured as logarithm via the acquired energy.~~

~~3. Method according to claim 1 or 2, whereby the global difference between the average word volume of a plurality of segmented words and the average silence volume of a plurality of segmented pauses is calculated, and the threshold is defined on the basis of the global difference.~~

~~4. Method according to claim 3, whereby the threshold is equated with the global difference.~~



averaged silence volume is employed in the determination of the difference

( $\Delta$ ):

13. Method according to one of the preceding claims, whereby an n-best list is prepared and a difference ( $\Delta$ ) between the average word volume ( $W_o$  level) of the individual spoken words and the average silence volume ( $S_i$  level) is allocated to each word of the n-best list, and the word to be inserted into the text from the n-best list is determined according to the criterion of this difference ( $\Delta$ ) between the average word volume ( $W_o$  level) of the individual spoken words and the average silence volume ( $S_i$  level):

14. Arrangement for speech recognition, comprising a processor unit that is configured such that

- a) words and pauses in the speech are determined on the basis of word boundaries;
- b) an average silence volume ( $S_i$  level) during the pauses is determined;
- c) an average word volume ( $W_o$  level) for the words is determined;
- d) a difference ( $\Delta$ ) is determined between the average word volume ( $W_o$  level) and the average silence volume ( $S_i$  level);
- e) speech is recognized when the difference ( $\Delta$ ) between the average word volume ( $W_o$  level) and the average silence volume ( $S_i$  level) is greater than a predetermined threshold ( $S$ );
- f) otherwise, no recognition of the speech is implemented.

Abstract

Method and Arrangement for Speech Recognition

